

APPLICATION FOR A UNITED STATES PATENT
UNITED STATES PATENT AND TRADEMARK OFFICE
(CASE No. 98,675-F)

Title: METHODS AND SYSTEM FOR EFFICIENT COLLECTION AND STORAGE OF EXPERIMENTAL DATA

5 **Inventors:** Keith S. Boyce, a citizen of the United States and the a resident of Wexford, Pennsylvania;

 Brian K. McKenna, a citizen of the United States and a resident of Pittsburgh, Pennsylvania;

10

 Phillip W. Glick, a citizen of the United States and a resident of Cheswick, Pennsylvania; and

15

 R. Terry Dunlay, a citizen of the United States and a resident of Plum Boro, Pennsylvania.

20 **Assignee:** Cellomics, Inc.
 635 William Pitt Way
 Pittsburgh, Pennsylvania 15328

CROSS REFERENCES TO RELATED APPLICATIONS

This application claims priority from U.S. Provisional Applications No.

60/108,291, filed on November 13, 1998, 60/110,643, filed on December 1, 1998,

- 5 60/140,240, filed on June 21, 1999, 60/142,375, filed on July 6, 1999, and 60/142,646
filed on July 6, 1999.

FIELD OF THE INVENTION

- 10 This invention relates to collecting and storing experimental data. More
specifically, it relates to methods and system for efficient collection and storage of
experimental data from automated feature-rich, high-throughput experimental data
collection systems.

BACKGROUND OF THE INVENTION

- 15 Historically, the discovery and development of new drugs has been an expensive,
time consuming and inefficient process. With estimated costs of bringing a single drug to
market requiring an investment of approximately 8 to 12 years and approximately \$350
to \$500 million, the pharmaceutical research and development market is in need of new
20 technologies that can streamline the drug discovery process. Companies in the
pharmaceutical research and development market are under fierce pressure to shorten
research and development cycles for developing new drugs, while at the same time, novel
drug discovery screening instrumentation technologies are being deployed, producing a
huge amount of experimental data.

Innovations in automated screening systems for biological and other research are capable of generating enormous amounts of data. The massive volumes of feature-rich data being generated by these systems and the effective management and use of information from the data has created a number of very challenging problems. As is
5 known in the art, "feature-rich" data includes data wherein one or more individual features of an object of interest (e.g., a cell) can be collected. To fully exploit the potential of data from high-volume data generating screening instrumentation, there is a need for new informatic and bioinformatic tools.

Identification, selection, validation and screening of new drug compounds is often
10 completed at a nucleotide level using sequences of Deoxyribonucleic Acid ("DNA"), Ribonucleic Acid ("RNA") or other nucleotides. "Genes" are regions of DNA, and "proteins" are the products of genes. The existence and concentration of protein molecules typically help determine if a gene is "expressed" or "repressed" in a given situation. Responses of genes to natural and artificial compounds are typically used to
15 improve existing drugs, and develop new drugs. However, it is often more appropriate to determine the effect of a new compound on a cellular level instead of a nucleotide level.

Cells are the basic units of life and integrate information from DNA, RNA, proteins, metabolites, ions and other cellular components. New compounds that may look promising at a nucleotide level may be toxic at a cellular level. Florescence-based
20 reagents can be applied to cells to determine ion concentrations, membrane potentials, enzyme activities, gene expression, as well as the presence of metabolites, proteins, lipids, carbohydrates, and other cellular components.

There are two types of cell screening methods that are typically used: (1) fixed cell screening; and (2) live cell screening. For fixed cell screening, initially living cells are treated with experimental compounds being tested. No environmental control of the cells is provided after application of a desired compound and the cells may die during screening. Live cell screening requires environmental control of the cells (e.g., temperature, humidity, gases, etc.) after application of a desired compound, and the cells are kept alive during screening. Fixed cell assays allow spatial measurements to be obtained, but only at one point in time. Live cell assays allow both spatial and temporal measurements to be obtained.

The spatial and temporal frequency of chemical and molecular information present within cells makes it possible to extract feature-rich cell information from populations of cells. For example, multiple molecular and biochemical interactions, cell kinetics, changes in sub-cellular distributions, changes in cellular morphology, changes in individual cell subtypes in mixed populations, changes and sub-cellular molecular activity, changes in cell communication, and other types of cell information can be obtained.

The types of biochemical and molecular cell-based assays now accessible through fluorescence-based reagents is expanding rapidly. The need for automatically extracting additional information from a growing list of cell-based assays has allowed automated platforms for feature-rich assay screening of cells to be developed. For example, the ArrayScan System by Cellomics, Inc. of Pittsburgh, Pennsylvania, is one such feature-

rich cell screening system. Cell based systems such as FLIPR, by Molecular Devices, Inc. of Sunnyvale, California, FMAT, of PE Biosystems of Foster City, California, ViewLux by EG&G Wallac, now a subsidiary of Perkin-Elmer Life Sciences of Gaithersburg, Maryland, and others also generate large amounts of data and photographic
5 images that would benefit from efficient data management solutions. Photographic images are typically collected using a digital camera. A single photographic image may take up as much as 512 Kilobytes ("KB") or more of storage space as is explained below. Collecting and storing a large number of photographic images adds to the data problems encountered when using high throughput systems. For more information on fluorescence
10 based systems, see "Bright ideas for high-throughput screening – One-step fluorescence HTS assays are getting faster, cheaper, smaller and more sensitive," by Randy Wedin, Modern Drug Discovery, Vol. 2(3), pp. 61-71, May/June 1999.

Such automated feature-rich cell screening systems and other systems known in the art typically include microplate scanning hardware, fluorescence excitation of cells,
15 fluorescence captive emission optics, a photographic microscopic with a camera, data collection, data storage and data display capabilities. For more information on feature-rich cell screening see "High content fluorescence-based screening," by Kenneth A. Giuliano, et al., Journal of Biomolecular Screening, Vol. 2, No. 4, pp. 249-259, Winter 1997, ISSN 1087-0571, "PTH receptor internalization," Bruce R. Conway, et al., Journal
20 of Biomolecular Screening, Vol. 4, No. 2, pp. 75-68, April 1999, ISSN 1087-0571, "Fluorescent-protein biosensors: new tools for drug discovery," Kenneth A. Giuliano and D. Lansing Taylor, Trends in Biotechnology, ("TIBTECH"), Vol. 16, No. 3, pp. 99-146, March 1998, ISSN 0167-7799, all of which are incorporated by reference.

An automated feature-rich cell screening system typically automatically scans a microplate plate with multiple wells and acquires multi-color fluorescence data of cells at one or more instances of time at a pre-determined spatial resolution. Automated feature-rich cell screen systems typically support multiple channels of fluorescence to collect multi-color fluorescence data at different wavelengths and may also provide the ability to collect cell feature information on a cell-by-cell basis including such features as the size and shape of cells and sub-cellular measurements of organelles within a cell.

The collection of data from high throughput screening systems typically produces a very large quantity of data and presents a number of bioinformatics problems. As is known in the art, "bioinformatic" techniques are used to address problems related to the collection, processing, storage, retrieval and analysis of biological information including cellular information. Bioinformatics is defined as the systematic development and application of information technologies and data processing techniques for collecting, analyzing and displaying data obtained by experiments, modeling, database searching, and instrumentation to make observations about biological processes.

The need for efficient data management is not limited to feature-rich cell screening systems or to cell based arrays. Virtually any instrument that runs High Throughput Screening ("HTS") assays also generate large amounts of data. For example, with the growing use of other data collection techniques such as DNA arrays, bio-chips, microscopy, micro-arrays, gel analysis, the amount of data collected, including photographic image data is also growing exponentially. As is known in the art, a "bio-chip" is a stratum with hundreds or thousands of absorbent micro-gels fixed to its surface.

A single bio-chip may contain 10,000 or more micro-gels. When performing an assay

test, each micro-gel on a bio-chip is like a micro-test tube or a well in a microplate. A bio-chip provides a medium for analyzing known and unknown biological (e.g., nucleotides, cells, etc.) samples in an automated, high-throughput screening system.

Although a wide variety of data collection techniques can be used, cell-based high throughput screening systems are used as an example to illustrate some of the associated data management problems encountered by virtually all high throughput screening systems. One problem with collecting feature-rich cell data is that a microplate plate used for feature-rich screening typically includes 96 to 1536 individual wells. As is known in the art, a "microplate" is a flat, shallow dish that stores multiple samples for analysis. A "well" is a small area in a microplate used to contain an individual sample for analysis. Each well may be divided into multiple fields. A "field" is a sub-region of a well that represents a field of vision (i.e., a zoom level) for a photographic microscope. Each well is typically divided into one to sixteen fields. Each field typically will have between one and six photographic images taken of it, each using a different light filter to capture a different wavelength of light for a different fluorescence response for desired cell components. In each field, a pre-determined number of cells are selected to analyze. The number of cells will vary (e.g., between ten and one hundred). For each cell, multiple cell features are collected. The cell features may include features such as size, shape, etc. of a cell. Thus, a very large amount of data is typically collected for just one well on a single microplate.

From a data volume perspective, the data to be saved for a well can be estimated by number of cell feature records collected and the number of images collected. The number of images collected can be typically estimated by: (number of wells x number of

fields x images per field). The current size of an image file is approximately 512 Kilobytes ("KB") of uncompressed data. As is known in the art, a byte is 8-bits of data. The number of cell feature records can typically be estimated by: (number of wells x number of fields x cells per field x features per cell). Data collected from multiple wells
5 on a microplate is typically formatted and stored on a computer system. The collected data is stored in format that can be used for visual presentation software, and allow for data mining and archiving using bioinformatic techniques.

For example, in a typical scenario, scanning one low density microplate with 96 wells, using four fields per well, three images per field and an image size of 512 Kbytes
10 per image, generates about 1,152 images and about 576 megabytes ("MB") of image data (i.e., $(96 \times 4 \times 3 \times 512 \times (1 \text{ KB} = 1024 \text{ bytes})) / (1 \text{ MB} = (1024 \text{ bytes} \times 1024 \text{ bytes})) = 576 \text{ MB}$). As is known in the art, a megabyte is 2^{20} or 1,048,576 bytes and is commonly interpreted as "one million bytes."

If one hundred cells per field are selected with ten features per cell calculated,
15 such a scan also generates $(96 \times 4 \times 100 \times 10) = 288,000$ cell feature records, whose data size varies with the amount of cell features collected. This results in about 12,000 MB of data being generated per day and about 60,000 MB per week, scanning the 96 well microplates twenty hours a day, five days a week.

In a high data volume scenario based on a current generation of feature-rich cell
20 screening systems, scanning one high-density microplate with 384 wells, using sixteen fields per well, four images per field, 100 cells per field, ten features per cell, and 512 KB per image, generates about 24,576 images or about 12,288 MB of image data and about 6,144,000 cell feature records. This results in about 14,400 MB of data being generated

per day and about 100,800 MB per week, scanning the 384 well microplates twenty-four hours a day, seven days a week.

Since multiple microplates can be scanned in parallel, and multiple automated feature-rich cell screening systems can operate 24 hours a day, seven days a week, and 365 days a year, the experimental data collected may easily exceed physical storage limits for a typical computer network. For example, disk storage on a typical computer network may be in the range from about ten gigabytes ("GB") to about one-hundred GB of data storage. As is known in the art, a gigabyte is 2^{30} bytes, or 1024 MB and is commonly interpreted as "one billion bytes."

The data storage requirements for using automated feature-rich cell screening on a conventional computer network used on a continuous basis could easily exceed a terabyte ("TB") of storage space, which is extremely expensive based on current data storage technologies. As is known in the art, one terabyte equals 2^{40} bytes, and is commonly interpreted as "one trillion bytes." Thus, collecting and storing data from an automated feature-rich cell screening system may severely impact the operation and storage of a conventional computer network.

Another problem with feature-rich cell screening systems is even though a massive amount of cell data is collected, only a very small percentage of the total cell feature data and image data collected will ever be used for direct visual display.

Nevertheless, to gather statistically relevant information about a new compound all of the cell data generated, is typically stored on a local hard disk and available for analysis.

This may also severely impact a local hard disk storage.

Yet another problem is that microplate scan results information for one microplate can easily exceed about 1,000 database records per plate, and cell feature data and image data can easily exceed about 6,000,000 database records per plate. Most conventional databases used on personal computers can not easily store and manipulate such a large number of data records. In addition, waiting relatively long periods of time to open such a large database on a conventional computer personal computer to query and/or display data may severely affect the performance of a network and may quickly lead to user frustration or user dissatisfaction.

Thus, it is desirable to provide a data storage system that can be used for feature-rich screening on a continuous basis. The data storage system should provide a flexible and scalable repository of cell data that can be easily managed and allows data to be analyzed, manipulated and archived.

SUMMARY OF THE INVENTION

In accordance with preferred embodiments of the present invention, some of the problems associated with collecting and storing feature-rich experimental data are overcome. Methods and system for efficient collection and storage of experimental data
5 is provided.

One aspect of the present invention includes a method for collecting experimental data. The method includes collecting image and feature data from desired sub-containers within a container. The image and feature data is stored in multiple image and feature databases. Summary data calculated for the desired sub-containers and the container are
10 stored in sub-container and container databases.

Another aspect of the present invention includes a method for storing experimental data on a computer system. The method includes collecting image data and feature data from desired sub-containers in a container. The image and feature data is stored in multiple third databases comprising multiple database tables. Summary data
15 calculated for desired sub-containers and the container is stored in a second database comprising multiple database tables. A first database is created that is a "pass-through" database. The first database includes a pass-through database table with links to the second database and links to the multiple third databases, but does not include any data collected from the container.

20 Another aspect of the present invention includes a method for spooling experimental data off devices that collect the data to a number of different remote storage locations. Links in a pass-through database table in a first database are updated to reflect the new locations of second database and multiple third databases.

Another aspect of the present invention includes a method for hierarchical management of experimental data. A pre-determined storage removal policy is applied to database files in a database. If any database files match the pre-determined storage removal policy, the database files are copied into a layer in a multi-layered hierarchical storage management system. The original database files are replaced with placeholder files that include a link to the original database files in the layer in the hierarchical storage management system.

Another aspect of the invention includes presenting the experimental data from a display application on a computer. The data presented by the display application is obtained from multiple databases obtained from multiple locations remote to the computer. The data displayed appears to be obtained from databases on local storage on the computer instead of from the remote locations.

Another aspect of the invention includes a data storage system that provides virtually unlimited amounts of "virtual" disk space for data storage at multiple local and remote storage locations for storing experimental data that is collected.

These methods and system may allow experimental data from high-throughput data collection systems to be efficiently collected, stored, managed and displayed. For example, the methods and system can be used for, but is not limited to, storing managing and displaying cell image data and cell feature data collected from microplates including multiple wells or bio-chips including multiple micro-gels in which an experimental compound has been applied to a population of cells.

The methods and system may provide a flexible and scalable repository of experimental data that can be easily managed and allows the data to be analyzed,

manipulated and archived. The methods and system may improve the identification, selection, validation and screening of new experimental compounds (e.g., drug compounds). The methods and system may also be used to provide new bioinformatic techniques used to make observations about experimental data.

- 5 The foregoing and other features and advantages of preferred embodiments of the present invention will be more readily apparent from the following detailed description. The detailed description proceeds with references to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the present invention are described with reference to
5 the following drawings, wherein:

FIG. 1A is a block diagram illustrating an exemplary experimental data storage system;

FIG. 1B is a block diagram illustrating an exemplary experimental data storage system;

10 FIG. 2 is a block diagram illustrating an exemplary array scan module architecture;

FIGS. 3A and 3B are a flow diagram illustrating a method for collecting experimental data;

FIG. 4 is a flow diagram illustrating a method for storing experimental data;

15 FIG. 5 is a block diagram illustrating an exemplary database system for the method of FIG. 4;

FIG. 6 is a block diagram illustrating an exemplary database table layout in an application database of FIG. 5;

20 FIG. 7 is a block diagram illustrating an exemplary database tables in a system database of FIG. 5;

FIG. 8 is a block diagram illustrating an exemplary database tables in an image and feature database of FIG. 5;

FIG. 9 is a flow diagram illustrating a method for spooling experimental data;

FIG. 10 is a flow diagram illustrating a method for hierarchical management
experimental data;

FIG. 11 is a flow diagram illustrating a method for presenting experimental data;
and

5 FIG. 12 is a block diagram illustrating a screen display for graphically displaying
experimental data.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Exemplary data storage system

FIG. 1A illustrates an exemplary data storage system 10 for preferred embodiments of the present invention. The exemplary data storage system 10 includes an analysis instrument 12, connected to a client computer 18, a shared database 24 and a data store archive 30 with a computer network 40. The analysis instrument 12 includes any scanning instrument capable of collecting feature-rich experimental data, such as nucleotide, cell or other experimental data, or any analysis instrument capable of analyzing feature-rich experimental data. As is known in the art, "feature-rich" data includes data wherein one or more individual features of an object of interest (e.g., a cell) can be collected. The client computer 18 is any conventional computer including a display application that is used to lead a scientist or lab technician through data analysis. The shared database 24 is a multi-user, multi-view relational database that stores data from the analysis instrument 12. The data archive 30 is used to provide virtually unlimited amounts of "virtual" disk space with a multi-layer hierarchical storage management system. The computer network 40 is any fast Local Area Network ("LAN") (e.g., capable of data rates of 100 Mega-bit per second or faster). However, the present invention is not limited to this embodiment and more or fewer, and equivalent types of components can also be used. Data storage system 10 can be used for virtually any system capable of collecting and/or analyzing feature-rich experimental data from biological and non-biological experiments.

FIG. 1B illustrates an exemplary data storage system 10' for one preferred embodiment of the present invention with specific components. However, the present invention is not limited to this one preferred embodiment, and more or fewer, and equivalent

types of components can also be used. The data storage system 10' includes one or more analysis instruments 12, 14, 16, for collecting and/or analyzing feature-rich experimental data, one or more data store client computers, 18, 20, 22, a shared database 24, a data store server 26, and a shared database file server 28. A data store archive 30 includes any of a disk
5 archive 32, an optical jukebox 34 or a tape drive 36. The data store archive 30 can be used to provide virtually unlimited amounts of "virtual" disk space with a multi-layer hierarchical storage management system without changing the design of any databases used to stored collected experimental data as is explained below. The data store archive 30 can be managed by an optional data archive server 38. Data storage system 10' components
10 are connected by a computer network 40. However, more or fewer data store components can also be used and the present invention is not limited to the data storage system 10' components illustrated in FIG. 1B.

In one exemplary preferred embodiment of the present invention, data storage system 10' includes the following specific components. However, the present invention
15 is not limited to these specific components and other similar or equivalent components may also be used. Analysis instruments 12, 14, 16, comprise a feature-rich array scanning system capable of collecting and/or analyzing experimental data such as cell experimental data from microplates, DNA arrays or other chip-based or bio-chip based arrays. Bio-chips include any of those provided by Motorola Corporation of Schaumburg, Illinois, Packard
20 Instrument, a subsidiary of Packard BioScience Co. of Meriden, Connecticut, Genometrix, Inc. of Woodlands, Texas, and others.

Analysis instruments 12, 14, 16 include any of those provided by Cellomics, Inc. of Pittsburgh, Pennsylvania, Aurora Biosciences Corporation of San Diego, California,

Molecular Devices, Inc. of Sunnyvale, California, PE Biosystems of Foster City, California, Perkin-Elmer Life Sciences of Gaithersburg, Maryland, and others. The one or more data store client computers, 18, 20, 22, are conventional personal computers that include a display application that provides a Graphical User Interface ("GUI") to a local
5 hard disk, the shared database 24, the data store server 26 and/or the data store archive 30. The GUI display application is used to lead a scientist or lab technician through standard analyses, and supports custom and query viewing capabilities. The display application GUI also supports data exported into standard desktop tools such as spreadsheets, graphics packages, and word processors.

10 The data store client computers 18, 20, 22 connect to the store server 26 through an Open Data Base Connectivity ("ODBC") connection over network 40. In one embodiment of the present invention, computer network 40 is a 100 Mega-bit ("Mbit") per second or faster Ethernet, Local Area Network ("LAN"). However, other types of LANs could also be used (e.g., optical or coaxial cable networks). In addition, the present
15 invention is not limited to these specific components and other similar components may also be used.

As is known in the art, OBDC is an interface providing a common language for applications to gain access to databases on a computer network. The store server 26 controls the storage based functions plus an underlying Database Management System
20 ("DBMS").

The shared database 24 is a multi-user, multi-view relational database that stores summary data from the one or more analysis instruments 12, 14, 16. The shared database 24 uses standard relational database tools and structures. The data store archive 30 is a

library of image and feature database files. The data store archive 30 uses Hierarchical Storage Management ("HSM") techniques to automatically manage disk space of analysis instruments 12, 14, 16 and the provide a multi-layer hierarchical storage management system. The HSM techniques are explained below.

5 An operating environment for components of the data storage system 10 and 10' for preferred embodiments of the present invention include a processing system with one or more high-speed Central Processing Unit(s) ("CPU") and a memory. In accordance with the practices of persons skilled in the art of computer programming, the present invention is described below with reference to acts and symbolic representations of
10 operations or instructions that are performed by the processing system, unless indicated otherwise. Such acts and operations or instructions are referred to as being "computer-executed" or "CPU executed."

 It will be appreciated that acts and symbolically represented operations or instructions include the manipulation of electrical signals by the CPU. An electrical
15 system represents data bits which cause a resulting transformation or reduction of the electrical signals, and the maintenance of data bits at memory locations in a memory system to thereby reconfigure or otherwise alter the CPU's operation, as well as other processing of signals. The memory locations where data bits are maintained are physical locations that have particular electrical, magnetic, optical, or organic properties
20 corresponding to the data bits.

 The data bits may also be maintained on a computer readable medium including magnetic disks, optical disks, organic memory, and any other volatile (e.g., Random Access Memory ("RAM")) or non-volatile (e.g., Read-Only Memory ("ROM")) mass

storage system readable by the CPU. The computer readable medium includes cooperating or interconnected computer readable medium, which exist exclusively on the processing system or be distributed among multiple interconnected processing systems that may be local or remote to the processing system.

5 **Array scan module architecture**

FIG. 2 is a block diagram illustrating an exemplary array scan module 42 architecture. The array scan module 42, such as one associated with analysis instrument 12, 14, 16 (FIG. 1B) includes software/hardware that is divided into four functional
10 groups or modules. However, more or fewer functional modules can also be used and the present invention is not limited to four functional modules. The Acquisition Module 44 controls a robotic microscope and digital camera, acquires images and sends the images to the Assay Module 46. The Assay Module 46 "reads" the images, creates graphic overlays, interprets the images collects feature data and returns the new images and
15 feature data extracted from the images back to the Acquisition Module 44. The Acquisition Module 44 passes the image and interpreted feature data to the Data Base Storage Module 48. The Data Base Storage Module 48 saves the image and feature information in a combination of image files and relational database records. The store clients 18, 20, 22 use the Data Base Storage Module 48 to access feature data and images
20 for presentation and data analysis by the Presentation Module 50. The Presentation Module 50 includes a display application with a GUI as was discussed above.

Collection of experimental data

FIGS. 3A and 3B are a flow diagram illustrating a Method 52 for collecting experimental data. In FIG. 3A at Step 54, a container with multiple sub-containers is

initialized using configuration information. At Step 56, the configuration information used for the container is stored in a container database. At Step 58, a loop is entered to repeat Steps 60, 62, 64, 66, 68, 70 and 72 for desired sub-containers in the container. At Step 60, a sub-container in the container is selected. In a preferred embodiment of the present invention, all of the sub-containers in a container are analyzed. In another embodiment of the present invention, less than all of the sub-containers in a container are analyzed. In such an embodiment, a user can select a desired sub-set of the sub-containers in a container for analysis. At Step 62, image data is collected from the sub-container. At Step 64, the image data is stored in an image database. At Step 66, feature data is collected from the image data.

In FIG. 3B at Step 68, the feature data is stored in a feature database. In one embodiment of the present invention, the image database and feature databases are combined into a single database comprising multiple tables including the image and feature data. In another embodiment of the present invention, the image database and feature databases are maintained as separate databases.

At Step 70, sub-container summary data is calculated. At Step 72, the sub-container summary data is stored in a sub-container database. In one embodiment of the present invention, the sub-container database and the container database are combined into a single database comprising multiple tables including the sub-container and container summary data. In another embodiment of the present invention, the sub-container and container databases are maintained as separate databases. The loop continues at Step 58 (FIG. 3A) until the desired sub-containers within a container have

been analyzed. After the desired sub-containers have been processed in the container, the loop at Step 58 ends.

At Step 74 of FIG. 3B, container summary data is calculated using sub-container summary data from the sub-container database. At Step 76, the container summary data
5 is stored in the container database.

In a general use of the invention, at Step 66 features from any imaging-based analysis system can be used. Given a digitized image including one or more objects (e.g., cells), there are typically two phases to analyzing an image and extracting feature data as feature measurements. The first phase is typically called "image segmentation" or
10 "object isolation," in which a desired object is isolated from the rest of the image. The second phase is typically called "feature extraction," wherein measurements of the objects are calculated. A "feature" is typically a function of one or more measurements, calculated so that it quantifies a significant characteristic of an object. Typical object measurements include size, shape, intensity, texture, location, and others.

15 For each measurement, several features are commonly used to reflect the measurement. The "size" of an object can be represented by its area, perimeter, boundary definition, length, width, etc. The "shape" of an object can be represented by its rectangularity (e.g., length and width aspect ratio), circularity (e.g., perimeter squared divided by area, bounding box, etc.), moment of inertia, differential chain code, Fourier
20 descriptors, etc. The "intensity" of an object can be represented by a summed average, maximum or minimum grey levels of pixels in an object, etc. The "texture" of an object quantifies a characteristic of grey-level variation within an object and can be represented by statistical features including standard deviation, variance, skewness, kurtosis and by

spectral and structural features, etc. The "location" of an object can be represented by an object's center of mass, horizontal and vertical extents, etc. with respect to a pre-determined grid system. For more information on digital image feature measurements, see: "Digital Image Processing," by Kenneth R. Castleman, Prentice-Hall, 1996, ISBN-0132114674, "Digital Image Processing: Principles and Applications," by G. A. Baxes, Wiley, 1994, ISBN-0471009490, "Digital Image Processing," by William K. Pratt, Wiley and Sons, 1991, ISBN-0471857661, or "The Image Processing Handbook - 2nd Edition," by John C. Russ, CRC Press, 1991, ISBN-0849325161, the contents of all of which are incorporated by reference.

10 In one exemplary preferred embodiment of the present invention, Method 52 is used to collect cell image data and cell feature data from wells in a "microplate." In another preferred embodiment of the present invention, Method 52 is used to collect cell image and cell feature data from micro-gels in a bio-chip. As is known in the art, a "microplate" is a flat, shallow dish that stores multiple samples for analysis and typically
15 includes 96 to 1536 individual wells. A "well" is a small area in a microplate used to contain an individual sample for analysis. Each well may be divided into multiple fields. A "field" is a sub-region of a well that represents a field of vision (i.e., a zoom level) for a photographic microscope. Each well is typically divided into one to sixteen fields. Each field typically will have between one and six photographic images taken of it, each
20 using a different light filter to capture a different wavelength of light for a different fluorescence response for desired cell components. However, the present invention is not limited to such an embodiment, and other containers (e.g., varieties of biological chips, such as DNA chips, micro-arrays, and other containers with multiple sub-containers),

sub-containers can also be used to collect image data and feature data from other than cells.

In an embodiment collecting cell data from wells in a microplate, at Step 54 a microplate with multiple wells is initialized using configuration information. At Step 56, the configuration information used for the microplate is stored in a microplate database. At Step 58, a loop is entered to repeat Steps 60, 62, 64, 66, 68, 70 and 72 for desired wells in the microplate. At Step 60, a well in the microplate is selected. At Step 62, cell image data is collected from the well. In one preferred embodiment of the present invention, the cell image data includes digital photographic images collected with a digital camera attached to a robotic microscope. However, other types of cameras can also be used and other types of image data can also be collected. At Step 64, the cell image data is stored in an image database. In another exemplary preferred embodiment of the present invention, the image database is a collection of individual image files stored in a binary format (e.g., Tagged Image File Format ("TIFF"), Device-Independent Bit map ("DIB") and others). The collection of individual image files may or may not be included in a formal database framework. The individual image files may exist as a collection of individual image files in specified directories that can be accessed from another database (e.g., a pass-through database).

At Step 66, cell feature data is collected from the cell image data. In one preferred embodiment of the present invention, Step 66 includes collecting any of the cell feature data illustrated in Table 1. However, other feature data and other cell feature can also be collected and the present invention is not limited to the cell feature data illustrated in Table 1. Virtually any feature data can be collected from the image data.

CELL SIZE
CELL SHAPE
CELL INTENSITY
CELL TEXTURE
CELL LOCATION
CELL AREA
CELL PERIMETER
CELL SHAPE FACTOR
CELL EQUIVALENT DIAMETER
CELL LENGTH
CELL WIDTH
CELL INTEGRATED FLUORESCENCE INTENSITY
CELL MEAN FLUORESCENCE INTENSITY
CELL VARIANCE
CELL SKEWNESS
CELL KURTOSIS
CELL MINIMUM FLUORESCENCE INTENSITY
CELL MAXIMUM FLUORESCENCE INTENSITY
CELL GEOMETRIC CENTER
CELL X-COORDINATE OF A GEOMETRIC CENTER
CELL Y-COORDINATE OF A GEOMETRIC CENTER

Table 1.

In FIG. 3B at Step 68, the cell feature data is stored in a cell feature database. In one embodiment of the present invention, the image database and cell feature databases are combined into a single database comprising multiple tables including the cell image
 5 and cell feature data. In another embodiment of the present invention, the image database (or image files) and feature databases are maintained as separate databases.

Returning to FIG. 3B at Step 70, well summary data is calculated using the image data and the feature data collected from the well. In one preferred embodiment of the present invention, the well summary data calculated at Step 72 includes calculating any
 10 of the well summary data illustrated in Table 2. However, the present invention is not limited to the well summary data illustrated in Table 2, and the other sub-containers and other sub-container summary data can also be calculated. Virtually any sub-container summary data can be calculated for desired sub-containers. In Table 2, a "SPOT" indicates a block of fluorescent response intensity as a measure of biological activity.

WELL CELL SIZES
WELL CELL SHAPES
WELL CELL INTENSITIES
WELL CELL TEXTURES
WELL CELL LOCATIONS
WELL NUCLEUS AREA
WELL SPOT COUNT
WELL AGGREGATE SPOT AREA
WELL AVERAGE SPOT AREA
WELL MINIMUM SPOT AREA
WELL MAXIMUM SPOT AREA
WELL AGGREGATE SPOT INTENSITY
WELL AVERAGE SPOT INTENSITY
WELL MINIMUM SPOT INTENSITY
WELL MAXIMUM SPOT INTENSITY
WELL NORMALIZED AVERAGE SPOT INTENSITY
WELL NORMALIZED SPOT COUNT
WELL NUMBER OF NUCLEI
WELL NUCLEUS AGGREGATE INTENSITY
WELL DYE AREA
WELL DYE AGGREGATE INTENSITY
WELL NUCLEUS INTENSITY
WELL CYTOPLASM INTENSITY
WELL DIFFERENCE BETWEEN NUCLEUS AND CYTOPLASM INTENSITY
WELL NUCLEUS BOX-FILL RATIO
WELL NUCLEUS PERIMETER SQUARED AREA
WELL NUCLEUS HEIGHT/WIDTH RATIO
WELL CELL COUNT

Table 2.

Returning to FIG. 3B at Step 72, the well summary data is stored in a well database. In one embodiment of the present invention, the well database and the microplate database are combined into a single database comprising multiple tables including the well and microplate data. In another embodiment of the present invention, the well and microplate databases are maintained as separate databases. Returning to FIG. 3A, the loop continues at Step 58 (FIG. 3A) until the desired sub-wells within a microplate have been analyzed.

After the desired wells have been processed in the microplate, the loop at Step 58 ends. At Step 74 of FIG. 3B, summary data is calculated using well summary data from

the microplate database. At Step 76, the microplate summary data is stored in the well database.

In one preferred embodiment of the present invention, the microplate summary data calculated at Step 74 includes calculating any of the microplate summary data
5 illustrated in Table 3. However, the present invention is not limited to the microplate summary data illustrated in Table 3, and other container and other container summary data can also be calculated. Virtually any container summary data can be calculated for a container. In Table 3, "MEAN" indicates a statistical mean and "STDEV" indicates a statistical standard deviation, known in the art, and a "SPOT" indicates a block of
10 fluorescent response intensity as a measure of biological activity.

MEAN SIZE OF CELLS MEAN SHAPES OF CELLS MEAN INTENSITY OF CELLS MEAN TEXTURE OF CELLS LOCATION OF CELLS NUMBER OF CELLS NUMBER OF VALID FIELDS STDEV NUCLEUS AREA MEAN SPOT COUNT STDEV SPOT COUNT MEAN AGGREGATE SPOT AREA STDEV AGGREGATE SPOT AREA MEAN AVERAGE SPOT AREA STDEV AVERAGE SPOT AREA MEAN NUCLEUS AREA MEAN NUCLEUS AGGREGATE INTENSITY STDEV AGGREGATE NUCLEUS INTENSITY MEAN DYE AREA STDEV DYE AREA MEAN DYE AGGREGATE INTENSITY STDEV AGGREGATE DYE INTENSITY MEAN MINIMUM SPOT AREA STDEV MINIMUM SPOT AREA MEAN MAXIMUM SPOT AREA STDEV MAXIMUM SPOT AREA MEAN AGGREGATE SPOT INTENSITY STDEV AGGREGATE SPOT INTENSITY
--

MCDONNELL BOEHREN
HUBERT & BENCHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

MEAN AVERAGE SPOT INTENSITY
STDEV AVERAGE SPOT INTENSITY
MEAN MINIMUM SPOT INTENSITY
STDEV MINIMUM SPOT INTENSITY
MEAN MAXIMUM SPOT INTENSITY
STDEV MAXIMUM SPOT INTENSITY
MEAN NORMALIZED AVERAGE SPOT INTENSITY
STDEV NORMALIZED AVERAGE SPOT INTENSITY
MEAN NORMALIZED SPOT COUNT
STDEV NORMALIZED SPOT COUNT
MEAN NUMBER OF NUCLEI
STDEV NUMBER OF NUCLEI
NUCLEI INTENSITIES
CYTOPLASM INTENSITIES
DIFFERENCE BETWEEN NUCLEI AND CYTOPLASM INTENSITIES
NUCLEI BOX-FILL RATIOS
NUCLEI PERIMETER SQUARED AREAS
NUCLEI HEIGHT/WIDTH RATIOS
WELL CELL COUNTS

Table 3.

In one exemplary preferred embodiment of the present invention, cell assays are created using selected entries from Tables 1-3. In a preferred embodiment of the present invention, a "cell assay" is a specific implementation of an image processing method

5 used to analyze images and return results related to biological processes being examined.

For more information on the image processing methods used in cell assays targeted to specific biological processes, see co-pending applications 09/031,217 and 09/352,171, assigned to the same Assignee as the present application, and incorporated herein by reference.

10 In one exemplary preferred embodiment of the present invention, the microplate and well databases are stored in a single database comprising multiple tables called "SYSTEM.MDB." The image and feature data for each well is stored in separate databases in the format "ID.MDB," where ID is a unique identifier for a particular scan.

However, the present invention is not limited to this implementation, and other types, and

15 more or fewer databases can also be used.

Storing experimental data

FIG. 4 is a flow diagram illustrating a Method 78 for storing collected experimental data. At Step 80, image data and feature data is collected from desired sub-containers in a container (e.g., with Method 52 of FIG. 3). At Step 82, a first database is
5 created. The first database includes links to other databases but does not include any data collected from the container. The first database is used as a "pass-through" database by a display application to view data collected from a container. At Step 84, a first entry is created in the first database linking the first database to a second database. The second database includes configuration data used to collect data from the container, summary
10 data for the container calculated from the desired sub-containers and summary data for the desired sub-containers in the container calculated from the image data and feature data. The information is organized in multiple database tables in the second database. At Step 86, multiple second entries are created in the first database linking the first database to multiple third databases. The multiple third databases include image data and feature
15 data collected from the desired sub-containers in the container. The data is organized in multiple database tables in the third database.

In one exemplary preferred embodiment of the present invention, at Step 80, image data and feature data is collected from desired wells in a microplate using Method 52 of FIG. 3. However, the present invention is not limited to using Method 52 to collect
20 experimental data and other methods can also be used. In addition, the present invention is not limited to collecting image data and feature data from wells in a microplate and other sub-containers and containers can also be used (e.g., bio-chips with multiple microgels).

At Step 82, an application database is created. In one exemplary preferred embodiment of the present invention, the application database includes links to other databases but does not include any data collected from the microplate. The application database is used by a display application to view data collected from a microplate. In
5 another embodiment of the present invention, the application database may include actual data.

FIG. 5 is a block diagram illustrating an exemplary database system 88 for Method 78 of FIG. 4. The database system 88 includes an application database 90, a system database 92 and multiple image and feature databases 94, 96, 98, 100. FIG. 5
10 illustrates only four image and feature databases numbered 1-N. However, the present invention is not limited to four image and features databases and typically hundreds or thousands of individual image and feature databases may actually be used. In addition the present invention is not limited to the databases or database names illustrated in FIG. 5 and more or fewer databases and other database names may also be used.

15 In one exemplary preferred embodiment of the present invention, the application database 90 is called "APP.MDB." However, other names can also be used for the application database in the database system and the present invention is not limited to the name described.

In one exemplary preferred embodiment of the present invention, a display
20 application used to display and analyze collected experimental does not access over a few thousand records at one time. This is because there is no need for evaluation of microplate detail data information (e.g., image or cell feature database data) across microplates. Summary microplate information is stored in microplate, well, microplate

feature and well feature summary tables to be compared across microplates. Detailed information about individual cells is accessed within the context of evaluating one microplate test. This allows a display application to make use of pass-through tables in the application database 90.

5 In a preferred embodiment of the present invention, the application database 90 does not contain any actual data, but is used as a "pass-through" database to other databases that do contain actual data. As is known in the art, a pass-through database includes links to other databases, but a pass-through database typically does not contain any actual database data. In such an embodiment, the application database 90 uses links
10 to the system database 92 and the multiple image and feature databases 94, 96, 98, 100 to pass-through data requests to the application database 90 to these databases. In another exemplary preferred embodiment of the present invention, the application database 90 may include some of the actual data collected, or summaries of actual data collected. In one exemplary preferred embodiment of the present invention, the application database
15 90 is a Microsoft Access database, a Microsoft Structured Query Language ("SQL") database or Microsoft SQL Server by Microsoft of Redmond, Washington. However, other databases such as Oracle databases by Oracle Corporation of Mountain View, California, could also be used for application database 90, and the present invention is not limited to Microsoft databases.

20 In another preferred embodiment of the present invention, a first pass-through database is not used at all. In such an embodiment, the first pass-through database is replaced by computer software that dynamically "directs" queries to/from the second and third databases without actually creating or using a first pass-through database.

FIG. 6 is a block diagram illustrating an exemplary database table layout 102 for the application database 90 of FIG. 5. The database table layout 102 of FIG. 6 includes a first pass-through database entry 104 linking the application database 90 to the system database 92. The database table layout also includes multiple second pass through database entries 106, 108, 110, 112 linking the application database to multiple image and feature databases 94, 96, 98, 100. However, more or fewer types of database entries can also be used in the application database, and the present invention is not limited to two types of pass-through databases entries. In another embodiment of the present invention, the application database 92 may also include experimental data (not illustrated in FIG. 6).

Returning to FIG. 4 at Step 84, a first entry is created in the application database 90 linking the application database 90 to a system database 92 (e.g., box 104, FIG. 6). The system database 92 includes configuration data used to collect data from a microplate, summary data for the microplate calculated from the desired wells and summary data for selected wells in the microplate calculated from the image data and feature data. This information is organized in multiple tables in the system database 92.

In one exemplary preferred embodiment of the present invention, the system database 92 is called "SYSTEM.MDB." However, other names could also be used and the present invention is not limited to this name. The system database 92 may also be linked to other databases including microplate configuration and microplate summary data and is used in a pass-through manner as was described above for the application database. In another exemplary preferred embodiment of the present invention, the

system database 92 is not linked to other databases, but instead includes actual microplate configuration and microplate summary data in multiple internal tables.

However, in either case, in one preferred embodiment of the present invention, the name of the system database 92 is not changed from microplate-to-microplate. In another preferred embodiment of the present invention, the name of the system database 92 is changed from microplate-to-microplate. A display application will refer to the system database 92 using its assigned name (e.g., SYSTEM.MDB) for microplate configuration and microplate summary data. Data stored in the system database 92 may be stored in linked databases so that the actual microplate container configuration and microplate summary data can be relocated without changing the display application accessing the system database 92. In addition the actual database engine could be changed to another database type, such as a Microsoft SQL Server or Oracle databases by Oracle, or others without modifying the display application accessing the system database 92.

FIG. 7 is a block diagram illustrating exemplary database tables 114 for the system database 92 of FIG. 5. The database table, 114 of FIG. 7 includes a plate table 116 that includes a list of plates being used. The plate table 116 is linked to a protocol table 118, a form factor table 122, a plate feature table 124 and a well table 126. The protocol table 118 includes protocol information. In a preferred embodiment of the present invention, a protocol specifies a series of system settings including a type of analysis instrument, an assay, dyes used to measure biological markers cell identification parameters and other parameters used to collect experimental data. An assay is described below. The form factor table 122 includes microplate layout geometry. For example, a

standard 96-well microplate includes 12 columns of wells labeled 1 through 12 and 8 rows of wells labeled A through H for a total of 96. The plate feature table 124 includes a mapping of features to microplates. The form factor table 122 is linked to the manufacturer table 120. The manufacturer table 120 includes a list microplate
5 manufactures and related microplate information. The well table 126 includes details in a well. In a preferred embodiment of the present invention, a well is a small area (e.g., a circular area) in a microplate used to contain cell samples for analysis.

The protocol table 118 is linked to a protocol assay parameters table 128. In a preferred embodiment of the present invention, an "assay" is a specific implementation of
10 an image processing method used to analyze images and return results related to biological processes being examined. The protocol assay parameters table 128 is linked to an assay parameters table 130. The assay parameters table 130 include parameters for an assay in use.

The protocol table 118 is also linked to a protocol channel table 132. Typically an
15 assay will have two or more channels. A "channel" is a specific configuration of optical filters and channel specific parameters and is used to acquire an image. In a typical assay, different fluorescent dyes are used to label different cell structures. The fluorescent dyes emit light at different wavelengths. Channels are used to acquire photographic images for different dye emission wavelengths. The protocol channel table
20 132 is linked to a protocol channel reject parameters table 134. The protocol channel reject parameters table 134 includes channel parameters used to reject images that do not meet the desired channel parameters.

The protocol table 118 is also linked to a protocol scan area table 136. The protocol scan area table 136 includes methods used to scan a well. The protocol scan area table 136 is linked to a system table 138. The system table 138 includes information configuration information and other information used to collect experimental data.

5 The well table 126 is linked to a well feature table 140. The well feature table 140 includes mapping of cell features to wells. The well feature table 140 is linked to a feature type table 142. The feature type table 142 includes a list of features (e.g., cell features) that will be collected. However, more or fewer tables can also be used, more or fewer links can be used to link the tables, and the present invention is not limited to the
10 tables described for the system database 92.

Returning to FIG. 4 Step 86, multiple second entries (e.g., boxes 106, 108, 110, 112 of FIG. 6) are created in the application database 92 linking the application database 92 to multiple image and feature databases 94, 96, 98, 100. The multiple image and feature databases include image data and feature data collected from the desired wells in
15 the microplate. The data is organized in multiple database tables in the image and feature databases.

In one exemplary preferred embodiment of the present invention, names of image and feature databases 94, 96, 98, 100 that contain the actual image and feature data are changed dynamically from microplate-to-microplate. Since the image and feature data
20 will include many individual databases, an individual image and feature database is created when a microplate record is created (e.g., in the plate table 116 (FIG. 7) in the system database 92 (FIG. 5)) and has a name that is created by taking a plate field value and adding ".MDB" to the end. (For example, a record in a plate table 116 with a field

identifier of "1234569803220001" will have it's data stored in a image and feature database with the name "1234569803220001.MDB"). However, other names can also be used for the image and feature databases and the present invention is not limited to the naming scheme using a field identifier from the plate table 116.

5 FIG. 8 is a block diagram illustrating exemplary database tables 144 for image and feature databases 94, 96, 98, 100 of FIG. 5. In one preferred embodiment of the present invention, the image and feature databases for a microplate include tables to hold image and feature data and a copy of the tables 116-142 (FIG. 7) excluding the manufacturer table 120 and the system table 138 used for the system database 92. In
10 another embodiment of the present invention, the image and feature databases 94, 96, 98, 100 include a copy or all of the tables 116-142 (FIG. 7). In another embodiment of the present invention, the image and feature databases 94, 96, 98, 100 do not include a copy of the tables 116-142 (FIG. 7) used for the system database 92. However, having a copy of the system database 92 tables in the image and feature databases allows individual
15 image and feature databases to be archived and copied to another data storage system for later review and thus aids analysis.

 The image and feature databases 94, 96, 98, 100, tables 144 include a well field table 146 for storing information about fields in a well. The well field table 146 is linked to a well feature table 148 that includes information a list of features that will be collected
20 from a well. The well field table 146 is also linked to a feature image table 150 that includes a list of images collected from a well and a cell table 152 that includes information to be collected about a cell. The cell table 152 is linked to a cell feature table 154 that includes a list of features that will be collected from a cell. However, more or

fewer tables can also be used, more or fewer links can be used to link the tables, and the present invention is not limited to the tables described for the image and feature databases.

Spooling experimental data

5 As was discussed above, the analysis instruments modules 12, 14, 16 generate a large amount of data including image data, feature data, and summary data for sub-containers and containers. The raw feature data values are stored as database files with multiple tables described above (e.g., FIG. 8). To prevent analysis instruments 12, 14, 16 and/or the store clients 18, 20, 22 from running out of file space, database files are
10 managed using a hierarchical data management system.

FIG. 9 is a flow diagram illustrating a Method 156 for spooling experimental data. At Step 158, a second database is copied from an analysis instrument to a shared database. The second database includes configuration data used to collect data from a container, summary data for the container calculated from one or more sub-containers in
15 the container and summary data for sub-containers in the container calculated from image data and feature data collected from desired sub-containers. The data in the second database is organized into one or more database tables. At Step 160, multiple third databases are copied to a shared database file server. The multiple third databases include image data and a feature data collected from the desired sub-containers in the
20 container. The data in the third database is organized into one or more database tables. At Step 162, a location of the second database and the one or more third databases is updated in a first database on the analysis instrument to reflect new storage locations for the second database on the shared database and one or more third databases on the shared

database file server. The first database includes links to the second database and the one or more third databases but does not include any data collected from the container. The first database is used by a display application to view data collected from a container.

In another preferred embodiment of the present invention, Method 156 further
5 comprises copying the first database from the analysis instruments 12, 14, 16 to a store client computers 18, 20, 22. Such an embodiment allows a display application on the store client computers 18, 20, 22 to view the data collected from the container using the first database copied to local storage on the client computers 18, 20, 22.

In another preferred embodiment of the present invention, Method 156 further
10 comprises locating the first database on the analysis instruments 12, 14, 16 from store client computers 18, 20, 22. Such an embodiment allows a display application on the store client computers 18, 20, 22 to view the data collected from the container at a remote location on the exemplary data storage system 10' from the store client computers 18, 20, 22.

15 The data collected is viewed from the display application on the store client computers 18, 20, 22 by retrieving container and sub-container data from the second database on the shared database 24 and image and feature data from the multiple third databases on the shared database file server 28.

In one exemplary preferred embodiment of the present invention, at Step 158, a
20 system database 92 (FIG. 5) is copied from an analysis instrument 12, 14, 16 to the shared database 24. The system database 92 includes configuration data used to collect data from a microplate, summary data for the microplate calculated from one or more wells in the microplate (e.g., Table 3) and summary data for wells in the microplate (e.g.,

Table 2) calculated from image data and feature data (e.g., Table 1) collected from desired wells as was described above. The data in the system database 92 is organized into one or more database tables (e.g., FIG. 7).

At Step 160, one or more image and feature databases 94, 96, 98, 100 are copied
5 to the shared database file server 28. The one or more image and feature databases 94, 96, 98, 100 include image data and a feature data collected from the desired wells in the microplate. The data in the one or more image and feature databases is organized into one or more database tables (e.g., FIG. 8).

At Step 162, a location of the system database 92 and the one or more image and
10 feature databases 94, 96, 98, 100 is updated in an application database 90 (FIG. 6) on the analysis instrument 12, 14, 16 to reflect new storage locations for the system database 92 on the store database 24 and one or more image and feature databases 94, 96, 98, 100 on the store archive 28.

In one preferred embodiment of the present invention, the application database 90
15 is a pass-through database that includes links (e.g., FIG. 6) to the system database 92 and the one or more image and feature databases 94, 96, 98, 100 but does not include any data collected from the microplate. In another embodiment of the present invention, the application database 90 includes data from the microplate. The application database 92 is used by a display application to view data collected from a microplate. However, the
20 present invention is not limited to this embodiment and other containers, sub-containers, (e.g., bio-chips with multiple micro-gels) and databases can also be used.

Hierarchical management of experimental data

FIG. 10 is a flow diagram illustrating a Method 164 for hierarchical management of experimental data. At Step 166, a hierarchical storage manager is initialized with a pre-determined storage removal policy. At Step 168, the hierarchical storage manager applies the pre-determined storage removal policy to database files in a database. At Step 5 170, a test is conducted to determine whether any database files on the database match the pre-determined storage removal policy. If any database files in the database match the pre-determined storage removal policy, at Step 172, the database files are copied from the database to a layer in a hierarchical store management system. At Step 174, database files in the database are replaced with placeholder files. The placeholder files include 10 links to the actual database files copied to the layer in the hierarchical store management system. If no database files in the database match the pre-determined storage removal policy, at Step 176, no database files are copied from the database to a layer in a hierarchical store management system.

In one exemplary preferred embodiment of the present invention, the pre- 15 determined storage removal policy includes one or more rules illustrated by Table 4. However, more or fewer storage removal policy rules can also be used and the present invention is not limited to storage removal policy rules illustrated in Table 4.

PERCENTAGE OF DISK SPACE AVAILABLE OR PERCENTAGE OF DISK SPACE USED. NUMBER OF FILES. DATE A FILE IS STORED. SIZE OF A FILE. NUMBER OF DAYS SINCE A FILE WAS LAST ACCESSED. FILE TYPE. FILE NAME.

Table 4.

Method 164 includes HSM steps that provide a method to allow on-line access to 20 virtually unlimited amounts of "virtual" disk space on data storage system 10'. The

virtual disk space is provided with a multi-layer hierarchical storage management system. The virtual disk space is provided without changing the layout of any database and is "invisible" to a user.

In one exemplary preferred embodiment of the present invention, the HSM steps
5 of Method 164 provide an archival method that implements a three-layer storage hierarchy including the disk archive 32, the optical jukebox 34 and the tape drive 36. However, more or fewer layers of storage can also be used and the present invention is not limited to HSM techniques with three-layer storage. Additional storage layers in the storage hierarchy are added as needed without changing the layout of any database or the
10 functionality of the hierarchical storage manager. The hierarchical storage manager can copy database files to layers in an N-layer storage hierarchy without modification.

In addition, virtually unlimited amounts of "virtual" disk space can be provided with a three-layer hierarchical storage management system by periodically removing re-writeable optical disks, from the optical jukebox 34 and tapes from the tape drive 36
15 when these storage mediums are filled with data. The re-writeable optical disks and tapes are stored in a data library for later access. In another preferred embodiment of the present invention, the data library is directly accessible from computer network 40.

In a preferred embodiment of the present invention, Method 164 supports at least two modes of database file archiving. However, more or fewer modes of database
20 archiving can also be used and the present invention is not limited to the two modes described.

In the first mode, the store server 26 retains database files on individual analysis instruments 12, 14, 16, where they were originally generated. The store server 26 uses

Method 164 to automatically manage the free space on the analysis instrument 12, 14, 16 disks to move files into a layer in the three-tiered storage management system. To the end user the files will appear to be in the same directories where they were originally stored. However, the files may actually be stored on the disk archive 32, the optical jukebox 34, or in a Digital Linear Tape ("DLT") 36 library.

In the second mode, the store server 26 spools database files from the analysis instruments 12, 14, 16, to the shared database 24 and the shared database file server 30 (e.g., using Method 156). The store server's 26 in turn manages database files on the shared database file server 30 using Method 164. In the second mode, the files may also be stored on the disk archive 32, the optical jukebox 34, or in a DLT 36 library.

Experimental data presentation

As was discussed above, an analysis instrument 12, 14, 16 can generate a huge amount of experimental data. To be useful, the experimental data has to be visually presented to a scientist or technician for analysis.

FIG. 11 is a flow diagram illustrating a Method 178 for presenting experimental data. At Step 180, a list including one or more containers is displayed using a first database from a display application on a computer. The containers include multiple sub-containers. Image data and feature data was collected from the one or more containers. The first database is a pass-through database including links to other databases with experimental data. At Step 182, a first selection input is received on the display application for a first container from the list. At Step 184, a second database is obtained for the first container from a first remote storage location. The first remote location is remote to the computer running the display application. The second database includes

configuration data used to collect data from the first container, summary data for the first container calculated from the sub-containers in the first container and summary data for desired sub-containers in the first container calculated from image data and feature data collected from desired sub-containers. At Step 186, a second selection input is received
5 on the display application for one or more sub-containers in the first container. At Step 188, multiple third databases are obtained from a second remote storage location. The multiple third databases include image data and feature data collected from the one or more sub-containers in the first container. At Step 190, a graphical display is created from the display application including container and sub-container data from the second
10 database, image data and feature data from the multiple third databases collected from the one or more sub-containers. Data displayed on the graphical display will appear to be obtained from local storage on the computer instead of the first remote storage location and the second remote storage location.

In one exemplary preferred embodiment of the present invention, Method 178 is
15 used for displaying experimental data collected from microplates with multiple wells. However, the present invention is not limited to this embodiment and can be used for other containers and sub-containers besides microplates with multiple wells (e.g., bio-chips with multiple micro-gels).

In such an exemplary embodiment at Step 180, a list including multiple
20 microplates is displayed from a display application on a computer. The microplates include multiple wells. Cell image data and cell feature data were collected from the multiple microplates. The display application uses an application database 90 to locate other databases, including experimental data.

In one preferred exemplary embodiment of the present invention, the application database 90 is located on the exemplary data storage system 10' at a location remote from the computer including the display application. The application database 90 is used from the computer including the display application without copying the application database
5 90 from a remote location on the exemplary data storage system 10'.

In another exemplary preferred embodiment of the present invention, the application database 90 is copied from a location on the exemplary data storage system 10' to local storage on the computer including the display application. In such an embodiment, the application database 90 is copied to, and exists on the computer
10 including the display application.

At Step 182, a first selection input is received on the display application for a first microplate from the list. At Step 184, a system database 92 is obtained for the first microplate from a first remote storage location. The first remote storage location is remote to the computer running the display application. The system database 92 includes
15 configuration data used to collect data from the first microplate summary data for the first microplate calculated from the wells in the first microplate and summary data for desired wells in the first microplate calculated from image data and feature data collected from desired wells.

At Step 186, a second selection input is received on the display application for
20 one or more wells in the first microplate. At Step 188, multiple image and feature databases 94, 96, 98, 100 are obtained from a second remote storage location. The multiple image and feature databases 94, 96, 98, 100 include image data and feature data collected from the one or more wells in the first microplate. At Step 190, a graphical

display is created from the display application including microplate and well summary data from the system database 92, image data and feature data from the multiple image and feature databases 94, 96, 98, 100 collected from the one or more wells. Data displayed on the graphical display appears to be obtained from local storage on the computer instead of the first remote storage location and the second remote storage location.

FIG. 12 is a block diagram illustrating an exemplary screen display 192 for visually displaying experimental data from a display application. The screen display 192 includes a display of multiple sub-containers 194 in a container 196. The container 194 includes 384 sub-containers (numbers 1-24 x letters A-P or $24 \times 16 = 384$). The screen display 192 also includes container summary data 198, sub-container summary data 200, image data 202, and feature data 204. The screen display 192 is capable of displaying the data in both graphical formats and textual formats depending on user preferences. A user can select his/her display preferences from menus created by the display application (Not illustrated in FIG. 12). Screen display 192 illustrates exemplary data for sub-container A-3 illustrated by the blacked sub-container 206 in the container 196. Experimental data collected from a container is visually presented to a scientist or lab technician for analysis using Method 178 and screen display 192 with a pass-through database with multiple links to multiple databases from multiple remote locations.

In one exemplary preferred embodiment of the present invention, a Store Application Programming Interface ("API") is provided to access and use the methods and system described herein. As is known in the art, an API is set of interface routines used by an application program to access a set of functions that perform a desired task.

In one specific exemplary preferred embodiment of the present invention, the store API is stored in a Dynamic Link Library ("DLL") used with the Windows 95/98/NT/2000 operating system by Microsoft. The DLL is called "mvPlateData.DLL"

However, the present invention is not limited to storing an API in a Window's DLL or
5 using the described name of the DLL and other methods and names can also be used to store and use the API. As is known in the art, a DLL is library that allows executable routines to be stored and to be loaded only when needed by an application. The Store API in a DLL is registered with the Window's "REGSVR32.EXE" application to make it available to other applications. The Store API provides an interface access to plate, well
10 image and cell feature information and provides a facility to enter desired well feature information that will be collected.

These methods and system described herein may allow experimental data from high-throughput data collection/analysis systems to be efficiently collected, stored, managed and displayed. The methods and system can be used for, but is not limited to
15 storing managing and displaying cell image data and cell feature data collected from microplates including multiple wells or bio-chips including multiple micro-gels in which an experimental compound has been applied to a population of cells. If bio-chips are used, any references to microplates herein, can be replaced with bio-chips, and references to wells in a microplate can be replaced with micro-gels on a bio-chip and used with the
20 methods and system described.

The methods and system may provide a flexible and scalable repository of cell data that can be easily managed and allows cell data to be analyzed, manipulated and archived. The methods and system may improve the identification, selection, validation

and screening of new experimental compounds which have been applied to populations of cells. The methods and system may also be used to provide new bioinformatic techniques used to make observations about cell data.

It should be understood that the programs, processes, methods and systems
5 described herein are not related or limited to any particular type of computer or network system (hardware or software), unless indicated otherwise. Various types of general purpose or specialized computer systems may be used with or perform operations in accordance with the teachings described herein.

In view of the wide variety of embodiments to which the principles of the present
10 invention can be applied, it should be understood that the illustrated embodiments are exemplary only, and should not be taken as limiting the scope of the present invention.

For example, the steps of the flow diagrams may be taken in sequences other than those described, and more or fewer elements may be used in the block diagrams. While various elements of the preferred embodiments have been described as being
15 implemented in software, in other embodiments in hardware or firmware implementations may alternatively be used, and vice-versa.

The claims should not be read as limited to the described order or elements unless stated to that effect. Therefore, all embodiments that come within the scope and spirit of the following claims and equivalents thereto are claimed as the invention.